Hurry Up and Wait: Differential Impacts of Congestion, Bottleneck Pressure, and Predictability on Patient Length of Stay

HARVARD | BUSINESS | SCHOOL

Hurry Up and Wait: Differential Impacts of Congestion, Bottleneck Pressure, and Predictability on Patient Length of Stay

Jillian Berry Jaeker
Anita L. Tucker

Working Paper

13-052

December 3, 2012

Hurry Up and Wait:

Differential Impacts of Congestion, Bottleneck Pressure, and Predictability on Patient Length of Stay

Jillian Berry Jaeker
Doctoral Candidate
jjaeker@hbs.edu

Anita L. Tucker,
Associate Professor
atucker@hbs.edu

Harvard Business School
Technology and Operations Management
Soldiers Field
Boston, MA
02163

ABSTRACT

High work load, from high inventory levels, impacts unit processing times, but prior operations management studies have found conflicting results regarding direction. Thus, it is difficult to predict inventory's effects on productivity a priori, inhibiting effective capacity management in high load systems. We categorize load into in-process inventory (congestion) and incoming inventory, decomposing the latter into its levels of bottleneck (BN) pressure and predictability, and quantify the magnitudes and directions of change on processing times. Using data from 283 hospitals, we find (1) high congestion increases a patient's hospital stay up to 28%, indicating inefficiencies from overloaded resources; (2) a patient stays up to 11.7% longer if there is a high load of incoming low BN pressure patients, consistent with the slowdown associated with "social loafing"; (3) a patient's stay is up to 10.2% shorter when there is a high incoming load of predictable patients, consistent with workload smoothing.

1. Introduction

Understanding the drivers of worker productivity is an important topic for managers and scholars. In particular, researchers have begun to examine the impact of increased inventory levels on productivity. We define productivity as the number of work units successfully processed in a given time period. It is determined, in part, by the average time it takes employees to complete the sets of tasks required for each "unit" of work. Traditional processing time models in operations management assume that the average processing time per unit, for the same type of units and a given production system, is driven solely by the task requirements, and therefore should be constant across multiple units of the same type. In other words, average processing times are assumed to be independent of state-specific contextual variables, such as the amount of work-in-process inventory, and behavioral factors, such as employees' responses to inventory levels.

However, recent empirical and analytical literature suggests that employees' processing times are influenced by inventory levels (Schultz, Juran et al. 1998; Oliva and Sterman 2001; Powell and Schultz 2004; Hopp, Iravani et al. 2007; KC and Terwiesch 2009; KC and Terwiesch 2012; Tan and Netessine 2012). The relationship between inventory and productivity is particularly strong in settings with high worker autonomy because workers have the ability to modify the set of work tasks performed on a unit, and the speed with which these tasks are performed (Hopp, Iravani et al. 2007). Furthermore, research suggests that work quality suffers as inventory in a system increases (Powell, Savin et al. 2011; Kuntz, Mennicken et al. 2012; Tan and Netessine 2012), which can further impact productivity by decreasing yield rates or increasing the amount of rework (KC and Terwiesch 2012). Failing to account for the impact of inventory on productivity can result in misalignment of labor (Green, Savin et al. 2011) and physical resources (Shapiro 1996; Green and Nguyen 2001). Thus, it is important to the study of productivity to better understand the link between inventory in the system and processing times.

We distinguish the inventory that is currently being processed in the system into more refined categories to better predict its effects. We use the term *congestion* to refer to the current inventory that is

being processed by the entire work area of interest ("processing area"). A particular worker's work-in-process inventory is her own workload and moves in parallel with the level of congestion in the work area. The upstream inventory waiting for processing by the processing area, is termed *incoming inventory*. Collectively, the incoming inventory and the congestion level comprise the system-wide inventory. We further decompose incoming inventory by the degree of *bottleneck (BN) pressure* and *predictability* associated with it, as determined by its source. This enables us to quantitatively determine the impact of different types of inventory streams, as well as their interactions, on processing times. By doing so, we are able to offer a priori predictions of the impact that inventory will have on processing times given its characteristics, such as its BN pressure and predictability.

This paper develops a set of hypotheses about the impact of inventory load on productivity which we test using patient-level data from more than 250 hospitals. We study hospitals because they have varying levels of incoming inventory (measured as patients awaiting an inpatient bed) and congestion (patients in the inpatient units of the hospital) (Green and Nguyen 2001). In addition, this setting has high worker autonomy (Eddy 1984; McLeod, Tamblyn et al. 1997), which enables workers to more easily adjust their processing times in response to inventory levels. Decreasing processing times without decreasing quality indicates increased productivity. In this setting, processing times refer to the length of stay (LOS) of patients in the hospital. Controlling for a patient's condition, we find that inpatient congestion is associated with an increase in the average LOS by up to 22.8% compared to when inpatient congestion is low. This slowing down is due to the negative effect of overloading a system's resources, as theorized by queuing theory. Average LOS is increased by up to 11.7% when there is a high number of incoming patients from low BN pressure locations, which is consistent with work avoidance due to social loafing. Conversely, LOS is reduced by up to 10.2% when there are a high number of predictable incoming patients compared to a high number of unpredictable incoming patients. This result is explained by employees' desire and ability to smooth their workload by completing current tasks in anticipation of an inflow of work. Finally, we find that the effects of both BN pressure and predictability are reduced when inpatient congestion is high, which is due to the reduced slack in their workload, making it difficult for employees to engage in either workload smoothing or social loafing. Our results suggest that adding capacity at the location of the inventory build-up (in our setting, the ED) may reduce productivity, and capacity should instead be added downstream (inpatient hospital units). Furthermore, we find, rather counter-intuitively, that in a setting with high worker discretion, more units may be processed in the same amount of time if there is additional slack capacity in resources, and the occupancy level is not maximized.

Our work contributes to the stream of operations management research on productivity firstly by decomposing the impacts of congestion, BN pressure and predictability on processing times. We link our

results to existing operations theory to explain the prior conflicting results of the impact of high load on worker productivity. This enables us to develop a priori predictions about the impact that high inventory levels will have on worker productivity. This impact depends on whether the inventory is currently being processed or, if it is incoming inventory, whether it has low BN pressure or high predictability. This allows us to provide guidance to managers on how to manage capacity to minimize the negative impacts from high loads. Second, we examine how the interaction of these three inventory characteristics further changes processing times. Third, we quantify these direct and interactive effects of inventory on processing times so that the relative impacts of congestion, predictability, bottleneck predictability, and their interactions can be compared.

2. Related Literature

In traditional operational models, the time to complete the work on a particular unit is assumed to be a random variable with a constant mean in which each unit of work is independent from each other unit (Dallery and Gershwin 1992). However, recent research has argued that the assumption of a constant mean is not always valid because workers can speed up or slow down when they have discretion over two key behaviors: how many tasks they perform for customers (Oliva and Sterman 2001; Hopp, Iravani et al. 2007; Batt, Terwiesch et al. 2012) and—even if they have no discretion over which tasks to perform—how long they take to perform a standard set of tasks (Schultz, Juran et al. 1998; KC and Terwiesch 2009). For example, in service settings, a worker can respond to her own workload—as implied by the number of people waiting for service from her—by either increasing or decreasing the amount of work she performs per customer, which impacts quality and the average processing time (Oliva and Sterman 2001; Hopp, Iravani et al. 2007; KC and Terwiesch 2009; Powell, Savin et al. 2011; KC and Terwiesch 2012; Kuntz, Mennicken et al. 2012; Tan and Netessine 2012). Even when the number of tasks is fixed, and the work is standardized, inventory levels have still been found to influence processing times as a worker can complete a specific set of tasks by working faster or slower (Schultz, Juran et al. 1998; Schultz, Juran et al. 1999).

While there is growing acceptance that high inventory load impacts processing times, the direction of the effect remains unclear, a priori. In the rest of this section we will present the relevant literature on increased inventory and its impact on processing times, making particular note of the roles of congestion, BN pressure and predictability.

We have defined congestion as the amount of current inventory being processed in the work area of interest, and it moves in parallel with individual workload. We will draw on prior research related to the topics of either congestion or individual employee workload to understand the processing time effects

related to increased processing area inventory. In settings such as law firms, consulting firms, restaurants, product development, scientific research, and hospitals, there are multiple customers or projects being processed at any one time by a given number of workers with a given level of resources. We know from Little's Law (1961) that given a set service rate (a common assumption), as work in process (WIP) inventory increases, the throughput time will increase at a constant and predictable rate. However, in more discretionary settings, the service rate may change in a nonlinear fashion as WIP inventory increases, and therefore the throughput time may be less predictable.

Empirical research has shown that congestion does have an impact on worker behavior, and in general, workers are more efficient when their own workload increases, until they become overwhelmed, at which point, quality and speed deteriorate (KC and Terwiesch 2009; Kuntz, Mennicken et al. 2012; Tan and Netessine 2012). For example, KC and Terwiesch (2009) found that as workload increases in a hospital unit, workers speed up their movements on individual tasks, but this increased speed can only last for a few hours before burn out sets in and fatigue results in more errors. Similarly, Tan and Netessine (2012) found in a restaurant setting that as workload increases (as measured by diners/waiter/hour), at first waiter performance and quality of service increases, but on average, once the number of diners per server increases around 34% above the mean, waiters become overwhelmed and speed up to reduce their own workload. While speeding up is generally a signal of efficiency, in this setting, speeding up can be a form of "cutting corners" (Oliva and Sterman 2001), which reduces service quality and leads to lower revenue. Using hospital data, Kuntz et al. (2012) found that as congestion on a hospital ward increases, quality, as measured by survival, also increases. However, like Tan and Netessine, Kuntz et al. found that once the amount of congestion passes a certain level (in this case around 92.5%), quality quickly drops, suggesting that workers become overwhelmed and either make mistakes or are unable to notice changes in patient status until it is too late. These empirical studies support the theory that worker behavior changes with increased system inventory in an inverted-U fashion, with workers initially increasing their efficiency with demand, but eventually decreasing their efficiency as demand gets very high because there is a limit to the number of concurrent customers that any given worker can handle while still maintaining quality. In a hospital setting, it has even been shown that when a nurse knows that his relative workload will be extremely high (in the negative slope portion of the inverted-u), he is more likely to call in absent to avoid being overwhelmed (Green, Savin et al. 2011). Tan and Netessine (2012) offer as an explanation for the inverted U-shaped relationship between worker performance and workload: workers smooth their workload such that they neither have idle time nor neglect some customers. Their study showed that when waiters are underworked, they find work-related tasks to occupy their time, but when overworked, they do not have the time to fully serve each diner and cut corners, negatively impacting quality and profitability.

While congestion accounts for the inventory in the processing area, there is another major category of inventory frequently studied in the field of productivity in operations: upstream, or incoming, inventory waiting to enter the processing area. Research suggests that two characteristics, BN pressure and predictability of this incoming inventory play a significant role in the effect of inventory on processing times for units currently being worked on in the processing area. There is a significant amount of research that examines the direct impact of high incoming inventory load on processing times, primarily in queuing theory and service settings. However, as the literature review below shows, the direction of the change in processing times associated with increased incoming inventory varies depending on the setting. We will argue that these perceived differences in processing time changes are attributable to the level of BN pressure and predictability in the presence of a high incoming inventory load.

Analytically, some models predict higher incoming queues of customers lead to longer processing times per customer, while other models predict a speeding up effect of longer queues. George and Harrison (2001) and Stidham and Weber (1989), among others, have analytically shown that as a queue increases in length (i.e.: high incoming inventory), and assuming a non-negative holding cost, the cost optimizing behavior is for worker processing speed to increase monotonically. However, in support of increased incoming inventory being associated with increased processing times, Ha (1998) finds that when there is no cost to the current customer (or worker) to attain a higher service quality by spending more time completing a transaction, too much time, compared to the system utility maximizing optimum, will be spent on the customer's tasks. This result is because there is a waiting cost for the customers in the queue who consequently have a longer wait before they receive service. Additionally, Hopp, Iravani et al. (2007) showed that under certain circumstances, when additional servers are brought in, lowering processing time pressure, customer processing times increase as workers improve service quality. These processing time increases are more likely to be maintained even as incoming inventory levels increase and longer queues form despite the increased service capacity, because there is less pressure for an individual worker to work faster and service rates are "sticky" (Oliva and Sterman 2001).

Meanwhile, empirical research shows similarly divergent results. On one hand, productivity decreases when there is greater inventory in a system as the increased inventory hides any discrepancies in processing speeds (Schultz, Juran et al. 1998; Schultz, Juran et al. 1999; Spear and Bowen 1999) and encourages "social loafing", which states that people put in less effort when it is more difficult to evaluate each individual's effort (Latané, Williams et al. 1979). As the number of customers waiting for service increases, there may be an increased incentive to "free ride" and let other workers take on the additional workload (Mas and Moretti 2007). On the other hand, higher inventory levels have been shown to cause workers to reduce processing times to eliminate long queues in service settings due to high costs of delay

(Oliva and Sterman 2001; KC and Terwiesch 2009; Batt, Terwiesch et al. 2012; KC and Terwiesch 2012; Tan and Netessine 2012).

Based on this research, we posit a possible explanation for why high incoming inventory load can have a multi-directional effect on processing times: BN pressure. Schultz, Juran et al. (1999) found that in a system with limited inventory, it is easier to spot the bottleneck, and people speed up to avoid being perceived as the one slowing down the production line (Powell and Schultz 2004). Furthermore, Mas and Moretti (2007), in their paper on the productivity of supermarket cashiers, showed that when queues were present and a slower cashier is in the line of vision of a fast cashier, the slower cashier speeds up, thus reducing his processing times and increasing productivity. They attributed this speeding up to "social pressure." In our paper, building off of Shultz, Juran et al. (1999), we term this BN pressure to signify that workers whose effort levels are easily observable will reduce their own processing times to avoid being perceived as the bottleneck step in the process with high incoming inventory load. Conversely, if workers' effort level is difficult to observe—which is akin to low BN pressure—they will slow down in the presence of a high incoming inventory load because of social loafing.

In addition to the effect of the BN pressure of the incoming inventory, we also consider the predictability of the incoming inventory. It has been shown that when demand is highly variable, which equates to low workload predictability, slack capacity is necessary to accommodate peak demand periods (Fisher and Ittner 1999; Armony and Gurvich 2010). However, because most firms need to minimize overhead costs, slack capacity may be an infeasible option for dealing with demand spikes, resulting in queues (Roberts, Frutos et al. 1999; Green 2002/2003), and increased processing times (Fisher and Ittner 1999). Research also suggests that predictability may lead to reduced processing times in setting with discretionary task completion times. Loch and Terwiesch (2005) theorize that predictability could incentivize a worker to incur the additional costs required to reduce processing times for current inventory because they know it will minimize costly congestion in the future. Other researchers have also proposed that workers should smooth their workflow by reducing processing times on current inventory to make way for known incoming inventory (Cachon and Fisher 2000). For example, Armony and Gurvich (2010) analytically showed in a call center setting that if there is high predicted future demand, the optimal policy is for workers to not upsell current customers, which results in shorter processing times than when upselling is performed, even if other workers are idle, in order to be available when future demand occurs.

All of the above research shares the assumption, implicitly or explicitly, that workers will vary their effort levels, and corresponding processing speeds, when faced with a high incoming inventory load. However, reducing processing time by either increasing worker speed or reducing the number of tasks performed for a customer incurs additional costs associated with increased effort and/or lower quality (Loch and Terwiesch 2005; KC and Terwiesch 2012). Workers will choose to incur these costs only if

the additional benefits, such as reduced queue length/wait time or reduced chance that a worker will have a heavy load in the future (Green, Savin et al. 2011; Tan and Netessine 2012), are greater than the additional cost. As any congestion, BN pressure, and predictability inventory increases, the relative costs and benefits of increased effort shift, leading to these inconsistent findings in processing time changes observed in the literature. To the best of our knowledge, no work has decomposed inventory into congestion, BN pressure, and predictability, and quantified how each affects processing times with high inventory loads.

3. Hospital processes, model, and hypotheses

We hypothesize that the effect of inventory load on processing times is determined by the type of inventory, which we characterize by the inventory's location (incoming inventory versus inventory currently being processed in the work area) and, for incoming inventory, the levels of BN pressure and predictability associated with it. To quantify how high loads of each type of inventory affects processing times, we will examine the change in expected LOS of inpatient patients in the hospital. The patients in the hospital as well as those being admitted should be interpreted as inventory from an operations perspective (though we will refer to them as patients, patient census, or patient load going forward). The number of inpatients in the hospital represents the level of congestion, and, indirectly, the average workload of hospital personnel. Similarly, patients waiting to be admitted to the hospital from the Emergency Department (ED) or the Post Anesthesia Care Unit (PACU) are considered incoming inventory, and depending on the path a patient takes to being admitted, he represents a different level of BN pressure and predictability.

3.1 Model

To develop our understanding of the effect of inventory on LOS, we first create a model of patients' expected LOS. Beginning here enables us to outline the drivers of LOS, and the mechanism through which inventory load can impact it. We model LOS as a function of patient level controls (e.g. age, sex, date of birth, medical condition), hospital level controls (e.g. hospital of treatment, average severity of patients treated at the hospital during the patient's stay), the service rate, $SR_{i,h,t}$, for patient i at hospital h at time t, of the hospital employees, the total amount of possible work that could be performed on patient i, W_i, and an error term, ε, representing the stochastic nature of treatment response. (For a complete list of controls see Appendix 1.) The patient and hospital controls provide a baseline amount of time that a patient must be in the hospital to recover for a given medical condition with a given set of patient characteristics. The amount of work performed on a patient is driven by the service rate, SR, of that patient's care givers on that day. Modeling this, we have, for patient i, in hospital, h, with an expected LOS,T:

$$LOS_{i,h} = \beta_1 Controls_i + \beta_2 Controls_h + \beta_3 \sum_{t=1}^{T} SR_{i,h,t} * \rho_{i,h,t} W_i + \varepsilon$$

,where $\rho_{i,h,t}$ is the portion of possible work that is performed for patient *i* at hospital *h* at time *t,* and β_n (for n=1,2,3) is a vector of coefficients. We do not explicitly have the service rate of workers in the hospitals nor the total amount of work to be performed, but we are only interested in the change in service rate and amount of work completed, which is the effect of inventory on processing times. Therefore, we model the service rates as a baseline service rate that is affected by the amount and type of patient inventory in the hospital, to get the following:

$$SR_{i,h,t} * \rho_{i,h,t} W_{i,t} = \alpha_1 + \alpha_2 I_h + \alpha_3 OCC_{h,t} + \alpha_4 SS_{h,t} + \alpha_5 ES_{h,t} + \alpha_6 EM_{h,t} + \eta$$

Where *OCC* is the occupancy of the hospital at time t, and is a measure of congestion, while *I* is a dummy variable equal to 1 if the hospital=*h,* to control for any differences across hospitals. *SS* is the number of scheduled surgical incoming admissions, which—as we will explain later—have high BN pressure and high predictability. *ES* is the number of emergency surgeries being admitted, which also have high BN pressure, but have low predictability compared with scheduled surgeries. *EM* is the number of incoming emergency medical admissions, which have both low BN pressure and low predictability. These three streams of incoming patients are the primary paths for being admitted to a hospital (see *Figure 1*); in our study context, there are too few scheduled medical patients, which would comprise the last remaining combination of pressure and predictability (low BN pressure and high predictability) for us to test this combination. It is the coefficients of *SS, ES*, and *EM* that we are most interested in, as these represent the role of the different inventory types on processing times. The reasoning behind the inventory levels affecting the service rate comes from the commonly used assumption in queuing theory that workers can decide their effort level, *e*, which manifests itself as the service rate in a hospital setting, and they determine this effort level by maximizing the benefits, *b,* compared to the costs, *c,* associated with this effort level, such that

$$\max_e \; b(e) - c(e)$$

As we will hypothesize, the above inventory measures have differing costs and benefits associated with increased effort levels. When inventory load change across the differing types, we hypothesize that the cost/benefit equation changes and results in the observed behavioral changes.

FIGURE 1 ABOUT HERE

3.2 Congestion

In our study, we use inpatient load (occupancy) in the hospital, as our measure of congestion. The greatest costs associated with hospital care are for equipment, building, and labor, which are fixed in the short term (Roberts, Frutos et al. 1999). Consequently, many hospitals try to leverage these costs with high

utilization of their assets, usually targeting an 85% occupancy rate (Green and Nguyen 2001), leading to times where there is high system level congestion. This increased congestion is associated with increased mental strain (Kuntz, Mennicken et al. 2012; Tan and Netessine 2012). As a result, workers have an incentive to reduce their workload by discharging a patient sooner than expected, either by working faster or omitting tasks entirely (KC and Terwiesch 2009; Tan and Netessine 2012).

There are competing factors, however, that contribute to a current patient staying longer in the hospital during times of high congestion. First, although a worker can work faster to meet the needs of all of her current customers, it is difficult for her to sustain this fast pace for an extended period of time (KC and Terwiesch 2009). Second, the increased patient inventory makes observing an individual's effort more difficult (Schultz, Juran et al. 1999),which may tempt a worker to relieve workload pressure by postponing work tasks which she could perform during her shift to the next shift, resulting in longer lengths of stay. Third, high congestion means that each worker cares for more patients concurrently. From queuing theory and Little's Law (1961), we know that if WIP, in this case patients, increases, and the output rate does not also increase, or only increases slightly, then the throughput time, or LOS, must increase. Finally, because the worker is caring for more patients concurrently, the fragmented work day has more mental setups, decreasing worker efficiency (Tucker and Spear 2006). We conclude that in situations with high congestion, the worker will not be able to increase her speed enough to compensate for the factors that push lengths of stay to be longer. Therefore, we hypothesize that high congestion is associated with increased processing time.

Hypothesis 1: High current patient inventory (congestion), holding all other inventory constant, is associated with increased LOS for current patients.

3.3 BN pressure and Predictability of Upstream Inventory

Congestion is associated with the level of current inventory, but often high load in the literature refers to the level of incoming, or upstream, inventory. In the hospital, the incoming inventory refers to the patients being admitted to the hospital. We hypothesize that how these incoming patients—SS, ES, and EM patients—affect the processing times of current patients depends on the characteristics of these incoming patients, in particular the BN pressure and predictability associated with them.

3.3.1 BN pressure

One characteristic of incoming inventory that we hypothesize impacts the processing time for current patients is BN pressure. During times of high incoming inventory load, if there is capacity to handle the extra inventory (in a hospital, this corresponds to empty inpatient beds), or as capacity becomes available (ie: patients are discharged), then the incoming inventory will move into the processing area (inpatient care units) until it is full. This movement of new inventory into the processing area can result in increased congestion and workload, which workers try to avoid (Green, Savin et al. 2011). Green, Savin

et al. (2011) showed that one way workers avoid this extra workload is through absenteeism. We hypothesize a second way: a service rate slow down for the current inventory when BN pressure is low, and effort is therefore difficult to observe, which makes the worker appear busy, and not the BN in the system, and thus reduces the probability that a worker will have to take on a new work unit.

In the hospital setting, incoming surgical patients result in high levels of BN pressure for workers caring for current inpatients. As previously mentioned, after surgery, patients are moved to the PACU and recover while waiting for an inpatient bed. In the event that a bed is unavailable when the patient has recovered from surgery, the patient can stay longer than medically necessary in the PACU (Ziser, Alkobi et al. 2002), though this is very costly due to the low nurse to patient ratio in the PACU (Waddle, Evers et al. 1998). Furthermore, if the PACU were to reach capacity and be unable to accept patients, the operating room would have to shut down, which has been estimated to result in lost profits of up to $5,000 per hour (Macario, Dexter et al. 2001). Therefore, there is a lot of pressure for the inpatient unit bed to be ready for an SS or ES patient. Further increasing the pressure, the physician performing the surgery is also frequently the one caring for the patient in the hospital (the "admitting physician") and as a result she can observe what beds are available in the hospital. Consequently, she can argue for admission for her patient when beds are available or occupied by patients who could be discharged.

In contrast, when a patient is admitted from an ED (and does not need surgery), there is very little BN pressure on workers caring for current inpatients. While the exact steps for admission for EM patients differs by hospital, in general the treating ED physician has to find a hospital physician who will accept an EM patient and be the admitting physician. Once an admitting physician is found, a bed is requested and prepared for the patient. If a bed is not immediately available, the EM patient is now officially under the care of the admitting physician, while still occupying a bed in the ED, in what is known in the hospital as "boarding". Although boarding has been shown to have negative consequences on patient outcomes (Chalfin, Trzeciak et al. 2007), it does not have the same high direct costs as delays in the PACU because the ED is more easily and cheaply able to add staff and hallway stretchers, creating swing capacity for the backlog of patients. Additionally, physicians in EDs often do not have the same visibility or relationships with the physicians and nurses in the hospital inpatient units, and therefore cannot observe the availability of beds or pressure these workers to increase their efforts. Therefore, if the inpatient units are becoming congested with all of the incoming patients, and more patients still need to be admitted from the ED, we propose that it can induce "social loafing" behavior in inpatient hospital nurses. Specially, a nurse may increase her processing time for her current patients, either by working slower or performing additional tasks, in order to avoid being assigned another patient because the ED physician cannot observe her effort and put pressure on her for being the bottleneck.

ES and EM patients only differ in the level of BN pressure they exert on workers caring for current inpatients. We predict that when there are a high number of incoming ES patients, the LOS of current patients will be lower than when there are a high number of incoming EM patients.

Hypothesis 2: A high load of incoming patients who induce low BN pressure is associated with higher LOSs for current patients than when there is a high load of incoming patients who induce high BN pressure.

3.3.2 Predictability

Another incoming inventory characteristic that we argue plays a significant role is the predictability of a new patient's admission. As we described in the previous section, workers try to avoid periods of high congestion (Green, Savin et al. 2011); smoothing their work flow avoids over- and under-utilization of resources while maximizing quality and/or profits (Schultz, Juran et al. 1999; Armony and Gurvich 2010; Tan and Netessine 2012). In a supply chain setting, it has been shown that when there is even a small increase in the predictability, and thus ability to smooth out work flow, there can be significant increases in profitability from matching supply and demand (Fisher and Raman 1996). Putting this in the context of the cost/benefit framework, exerting effort on current inventory to alleviate the pressure from future demand has cost, in terms of both effort and uncertainty. If a hospital worker is unsure of the exact amount or type of future demand, then it is challenging to accurately predict the number and types of beds needed for future patients. Therefore, nurse managers and physicians can prepare for potential future patients by discharging current patients earlier than they otherwise might, but doing so requires them to predict which patient beds are most needed and therefore which patients should be discharged. Alternatively, they can wait longer before discharging current patients (which has the additional benefit of ensuring that the current patient is fully recovered) to gain more certainty about which new patients will arrive on the unit. This strategy of postponing the discharge of current patients has the downside of causing delays for incoming patients if the incoming stream of patients is larger or arrives sooner than anticipated. Loch and Terwiesch (2005) frame this speed/accuracy tradeoff as a choice of "rush and be wrong" or "wait and be late". Higher levels of predictability in future demand reduces the cost of rushing. Therefore, when incoming inventory is highly predictable, we hypothesize that exerting extra effort at the present time to smooth out future workflow is less risky, and as a result, less costly.

In the hospital setting, incoming patients are more predictable when they have been scheduled more than 24 hours in advance, as is the case for SS patients. Scheduling at least a day in advance enables managers to know that a new patient will be admitted to their unit and therefore they can account for that patient in their staffing plan and assign the patient to a worker at the start of the shift. Similar to call center workers reducing current upselling when there is known future demand (Armony and Gurvich 2010), a hospital worker can increase her effort on current patients so that at least one patient can be

discharged, reducing her workload so that she is better able to handle the time consuming and mentally taxing task of a new admission (Matthews, Harvey et al. 2002). While this is more work in the short term and perhaps holds the small risk that a patient will be discharged too soon, it will smooth out workflow so that the worker does not experience the negative effects of congestion. Thus, the effects of scheduled incoming patients on current patients' length of stay will be greatest when there are a lot of scheduled patients and a current patient is close to discharge. The incoming patients with the highest level of predictability are the SS patients. These are patients who are scheduled for surgery at least 24 hours ahead of time, as opposed to emergency surgical patients ES. For both SS and ES patients, the patient comes to the hospital on the day of the surgery (with the ES patient coming via the ED), is operated on, goes to the PACU until he is stable enough to be moved, and then he is transported into an inpatient unit to recover before being discharged from the hospital. However, the only difference is that in the SS case, the nursing unit that will receive this patient after his PACU stay is informed of the pending arrival in advance, which enables the unit to plan for that arrival by reserving a bed and nurse for his care. Thus, we can compare patients who are essentially the same except for the level of predictability associated with their admission paths. We predict that incoming SS patients will reduce the LOS of current patients more than incoming ES patients.

Hypothesis 3: A high load of highly predictable incoming patients is associated with lower LOS for current patients than when there is a high load of less predictable incoming patients.

3.4 Interaction of Congestion, BN pressure, and Predictability

In addition to the roles of inventory congestion, predictability, and BN pressure on processing time, there is also the effect of their interactions. In particular, we examine how high congestion interacts with BN pressure and predictability. We hypothesize that when the load of both inpatients and incoming patients are high, processing times will change in an amount different from the sum of the impact from each individual type of busyness.

3.4.1 Congestion and BN pressure

We hypothesized that (1) congestion increases LOS of current patients and that (2) incoming inventory associated with low BN pressure increases LOS compared to high BN pressure patients. Combining these two predictions, we anticipate that congestion will have a greater impact on increasing LOSs for patients who face high BN pressure than those who face low BN pressure. This is because patients who face low BN pressure already have longer LOSs and therefore there is less opportunity for their LOS to be increased further by congestion. Conversely high BN pressure is associated with reduced processing times because workers do not want to be perceived as the bottleneck, even if they face a heavy future workload from taking on additional work more quickly rather than relying on social loafing (Mas and Moretti 2007). However, increased congestion makes effort levels, and subsequent bottlenecks, less observable

(Schultz, Juran et al. 1999). This inability to observe effort when congestion is high is equivalent to reducing the BN pressure. Furthermore, as congestion increases, a worker has less slack capacity to handle mentally taxing new patients (Matthews, Harvey et al. 2002). Consequently, we predict this worker will increase her processing time (such as by performing additional tests on the patient or delaying writing discharge orders) on current patients because she does not want any additional incoming patients, and she faces less pressure when she increases her processing times due to the lack of effort visibility when congestion is high. As a result, the relative benefits of incoming inventory with high BN pressure are reduced when accompanied by an increase in congestion.

Hypothesis 4: Congestion moderates the impact of BN pressure on length of stay such that the reduction in length of stay from BN pressure is lower in the presence of congestion.

3.4.2 Congestion and Predictability

Hypothesizing the interaction effect of congestion and predictability is more complicated since we have hypothesized that increased congestion will increase patient LOS while high levels of predictability of incoming patients will decrease patient LOS. Although these conditions produce opposite results on processing times, by more closely examining how each changes processing times, we can hypothesize about their interaction. In the case of congestion, processing times increase because of increased mental strain and work in process (Tucker and Spear 2006; KC and Terwiesch 2009; Kuntz, Mennicken et al. 2012; Tan and Netessine 2012). For incoming inventory with high predictability, we hypothesize that processing times decrease because workers are able to rush prior to a new customer arriving, and thus smooth their own workflows (Fisher and Raman 1996; Loch and Terwiesch 2005). However, the ability to rush is based on having enough slack capacity (Fisher and Ittner 1999). If congestion is high, a worker will not be able to rush as easily in her current work because her time is consumed with current WIP. For example, returning to Armony and Gurvich's (2010) call center setting, if congestion is high, upselling will never be attempted, so there is no flexibility to reducing upselling, and thus reduce processing times, since workers are already at maximum output. In our setting, when there is congestion among inpatients and high levels of predictable incoming patients, the LOS of current patients will still decrease compared to high congestion with a low load of incoming patients, but the change in LOS will be less than when there is no congestion. Since a worker is less able to rush, the processing times of current inpatients will still decrease when there are high levels of predictability in the incoming patients, but this impact will be attenuated when there is a high congestion compared to when there is a low congestion.

Hypothesis 5: Congestion moderates the impact of predictability on length of stay such that the reduction in length of stay from predictability is lower in the presence of congestion.

4. Data

Our data consists of patient level records for all Emergency Department (ED) and inpatient discharges in the state of California from December, 2007 to December, 2009 (Office of Statewide Health Planning and Development). We sum all patient admissions and subtract all patient discharges for all patients admitted in December, 2007 to determine the baseline number of patients in each hospital on January 1, 2008. We restrict our sample to only include patients admitted after December 31, 2007, and before November 30, 2009 since our data only has patients who were discharged by December 31, 2009, and does not include any patients admitted in December, but discharged on or after January 1, 2010, thus making it impossible for us to get an accurate census during that month. We limit our data to acute care hospitals with at least 3500 patients per year to ensure enough patients per day to calculate reasonable occupancy levels, and a 24 hour ED to ensure there are emergency admissions, resulting in a sample of 283 hospitals. Each admission or visit to the ED is its own record and includes the date admitted, date discharged, demographic information on the patient, hospital of care, type of care, diagnoses, major procedures, disposition, if an admission was scheduled, and diagnosis related group (DRG), among others. As a result, we know whether the patient was surgical or medical, and if the visit was scheduled or an emergency. We use the Centers for Medicare & Medicaid Services (CMS) weighting system (2010) to provide a proxy for severity for each patient based on his DRG. This severity score is the multiplier that CMS uses when paying hospitals and is proportional to the resources used for the patient. Since surgical patients tend to use more resources for the same severity level, we categorize the severity score as either surgical or medical. Lastly, we have dates for all major procedures performed, and the LOS of each patient in days. We use the CMS (2010) average LOS as the expected LOS associated with each DRG, allowing us to calculate an expected discharge date for each patient based on his date of admission.

To calculate the effect of increased inventory on LOS, we analyze how high inventory load on the day before a patient's expected discharge impacts that patient's LOS. We consider high load on the day before expected discharge because most discharge preparations begin on the day before expected discharge (Litvak 2010). If there is a slowdown due to congestion on the day before discharge, discharge preparations may be postponed, resulting in that patient's discharge being delayed beyond the expected day of discharge. Conversely, if there is a speed-up, all the discharge preparations may be completed early, resulting in that patient being discharged on the day before expected discharge. If there is no speed-up or slowdown, the patient is most likely to be discharged on the expected day of discharge.

The hospitals in our data vary greatly in the number of admissions, so the absolute level of congestion and incoming inventory can similarly vary. To be able to analyze the effects of patient inventory load across these heterogeneous hospitals, we have converted the absolute number of patients currently in the hospital, as well as the number of incoming SS, ES, and EM patients, for each date, into a percentage of

the maximum for each inventory category. Specifically, we calculate the maximum number of patients treated or admitted on weekdays for each quarter-year combination, and then do the same for weekends for each quarter-year combination. We separate out weekdays and weekends to account for what is commonly termed the "weekend effect" in hospitals, and we distinguish the quarters of the years to account for seasonality in demand. For example, to calculate the congestion for hospital "H" on Monday, June 9, 2008, we first find the maximum number of patients who were inpatient on a single weekday in hospital H in quarter 2 of 2008. We then divide the number of patients actually in the hospital on June 9 by that maximum.

Since we predict a patient's LOS based on inventory on the day before expected discharge, we need to know the day of expected discharge and the effect due to inventory load. To accomplish this we focus our analysis on one DRG – hip or knee replacement. We chose this DRG for three reasons. It is one of the top five most common DRGs; there is a relatively small variation in LOS across patients, reducing noise in our sample; and the average LOS is relatively long, which is necessary for us to detect day level changes in LOS. Note that while we focus on the LOS of hip or knee replacement patients, the congestion and incoming patient volumes include all DRGs because the incoming work and current inventory in the system is comprised of all patient types.

5. Econometric Specifications

The LOS of a patient can be thought of as a survival function, with discharge being equivalent to exiting the system (KC and Terwiesch 2012). Survival analysis allows us to predict a patient's likelihood of discharge on any given day based on an underlying hazard function scaled by the patient's and hospital's characteristics. Since our LOS is at the day level, we must use a discrete-time survival analysis. In the medical literature it is common to allow the baseline hazard to vary with time (KC and Terwiesch 2012), and since our data does not fit a known distribution very well, we include a variable for each day to account for this flexible hazard rate.

We follow Jenkins' (2004) approach and use the proportional hazard complementary log-log (cloglog) regression to model the probability of "failure" (e.g. discharge) at any given time (in our case, for each day). Cloglog is an appropriate model since it can be used with discrete and censored survival times, and it is the best estimator when observation times are discrete but the events occur continuously; in our study we have daily census, admission, and discharge data, but patients are admitted and discharged throughout the day. This gives us the following hazard for patient *i* at time *t* :

$$h_i(t) = 1 - \exp[-\gamma_i \exp(h_0(t))]$$

Where $h_0(t)$ is the baseline hazard on day *t* and

$$\gamma_i = \exp(\beta' X_i)$$

in which β is a vector of coefficients, and X is a vector of patient and hospital characteristics, restricting our analysis of LOS to patients with the DRG of hip or knee replacement/revision. We calculate this hazard rate for each patient for each day, $d=1,2,\ldots17$, which allows us to (1) observe changes in the probability of being discharged across multiple days (i.e. if the LOS is expected to increase, then the biggest changes in the hazard rate function will occur on days 4 and 5, while conversely if LOS is expected to decrease, then the biggest changes will be on days 3 and 4), and (2) calculate the expected LOS for each patient type, which involves knowing the hazard rate for each day, up to the maximum expected LOS. We consider day 17 to be the maximum expected LOS and censor patients with a LOS greater than 17 days (more than 3 times the standard deviation away from the mean LOS, and where the probability of occurring drops off dramatically); in survival analysis this is interpreted as having had no event before the end of the observation period of 17 days. Based on the authors' observations, patients with such a high LOS generally have some other unobservable medical or social condition, rather than the patient census, that is causing the high LOS.

In addition to the controls listed in Appendix 1, we include our variables of interest: hospital occupancy, scheduled surgical (SS) admits, emergency surgical (ES) admits, and emergency medical (EM) admits. In our study, the expected LOS for the DRG of interest, hip or knee replacement, as provided by CMS, is 4 days. Therefore, we look at the occupancy and inventory levels of interest on day three. We measure the effects on LOS from high inventory load only on those patients currently in the hospital with the DRG of interest. This technique enables us to observe how LOS differs at vary inventory loads across patient types. To control for the service rate when a patient enters, we also include the occupancy of the hospital and relative number of admits for each category of inventory (e.g. congestion, BN pressure, and predictability) on the day of admission. We do not control for the level of busyness on other days of stay because of the high level of correlation between adjacent days which would overspecify the model if included. In addition, we control for the severity of incoming patients on day 3 since hospitals prioritize patients based on severity, and may be more likely to discharge a less severe patient if a more severe patient needs a bed (KC and Terwiesch 2012). The complete list of variables for our analysis can be found in Appendix 1

Given the set of hazard functions, we then solve for the patient's survival function $S_i(t)$, which gives the chances of surviving past time t, and equals

$$S_i(t) = \exp\left\{\sum_{d=1}^{t} \ln[1 - h_i(d)]\right\}, where$$

$$S_i(0) = 1 \; \forall \; i$$

As with the hazard function, we calculate these survival functions for each patient for each day, *d=1,2,…,17*, to yield a survival curve. We can also find the expected survival time of the patient, or LOS, of a patient with a specific set of characteristics, and the relative effects of each of these characteristics, as

$$E(LOS) = \sum_{t=1}^{K} t * [S_i(t) - S_i(t-1)]$$

, where *K* is the expected maximum LOS in days.

We use Stata 12 for all of our analysis. The hazard functions are solved using maximum likelihood estimation on our inpatient hospital dataset, yielding us a model of defined coefficients for each hazard function. It should be noted that cloglog regression assumes proportional hazards, but since we believe that workers behave differently when any of the inventory load is high, we run multiple analyses corresponding to different ranges of busyness (e.g. the hospital is busy, incoming scheduled surgical patients are busy, no other incoming patients types are busy), as recommended by Hoetker (2007) for logistic regressions, then compare the resulting expected survival times, averaged across all hospitals. We are interested in the effects of congestion, SS, ES, and EM admission rates, and we compare the associated expected LOS of each to the others to give us the effects of congestion, predictability, BN pressure, and their interactions, in total running eight different survival analysis functions to yield eight sets of hazard functions, and the corresponding survival curves and expected LOSs. These eight survival functions come from congestion, SS, ES, or EM all being low (baseline function), and then four more functions with one of the four being high and the other three being low (main effects). The final three functions are to test the interaction effects and have high congestion while one of SS, ES, or EM are high and the other two are low. See Appendix 2 for the complete list of survival function analyses.

To run these survival functions, we have to define busyness for each variable of interest. We define busyness for inpatient hospital occupancy as any day with an occupancy level greater than 85% of the maximum value. This approach is consistent with prior research that generally defines high load in hospitals as somewhere between 85% and 93% of capacity (Green and Nguyen 2001; KC and Terwiesch 2012; Kuntz, Mennicken et al. 2012). We choose 85% as we are looking at an entire hospital's occupancy, not just one unit, which is often the measure for target occupancy. Furthermore, we want to be conservative in our estimates of busyness to ensure that pick up the effect and can maintain the proportional hazards assumption. Using 85% also allows us to have enough observations to calculate each hazard function with sufficient power. In our sample, 85% corresponds to the approximate midpoint of the distribution of hospital occupancy, thus ensuring a sufficient sample size. To be consistent, we similarly define the busyness levels for SS, ES, and EM daily admissions as the midpoint for each distribution, yielding busyness definitions for incoming patients as admission levels greater than 50% for SS, 46% for ES, and 62% for EM.

Following Hoetker's (2007) recommended approach, we use the coefficients defined by the cloglog regression in combination with a set of defined values for our control and independent variables. By putting these values into our model, we can come up with a numerical probability of survival for each day at each hospital, and from there the expected LOSs for each hospital, for our 8 functions of interest. Since we compare across survival functions, we want our control variables to have the same values in each analysis, and to represent a "typical" patient, so we set each control variable to its median/modal value (See Appendix 1 for values). We want to know how high levels of the different types of inventory load affect processing times and we need to separate out each type of inventory as cleanly as possible, particularly for the incoming inventory. Thus, to create conditions where the hospital is busy, we set the occupancy level to the maximum level (100%), while when the hospital is not busy we set the occupancy to 50%. In this case, these represent the 99th and 1st percentile of hospital occupancy, respectively, and thus the different ends of the load spectrum. For SS, ES, and EM admission levels, when not busy, we set them equal to 0 to ensure that we only get an effect from the incoming inventory type of interest, and when busy, we set them to 100%. We understand that these values provide us an upper bound, but think it provides the cleanest comparison of the inventory types, and we are not focused on the precise value of the effect of each type of inventory, but the direction and relative order of magnitude of the impact of each inventory type.

Using this method, we can calculate and compare the average expected LOSs of "hypothetical" patients with the eight different inventory profiles, one patient for each profile from each of the 283 hospitals. We then create an average expected LOS by averaging the 283 patients with the same profile across all hospitals. For hypothesis 1, we compare the average expected LOS of the hypothetical hip/knee replacement patients (n=283, one from each hospital) who experience high congestion (100% hospital occupancy) on the day before expected discharge to the same type of patients, but who do not experience high congestion on the day before expected discharge (50% hospital occupancy). To test Hypothesis 2, we compare the expected LOS of current knee/hip replacement patients who, on the day before expected discharge, are in a hospital that has a high incoming load of EM patients—which represents the scenario where there is low level of BN pressure—to patients who, on the day before expected discharge, are in a hospital that has high load of incoming ES patients—which is the scenario with a high level of BN pressure. By comparing these two groups, we are able to hold constant the impact on LOS from unscheduled, ED arrivals, and from increased incoming inventory in general, and only measure the effect of BN pressure. Similarly for H3 we compare a high load of incoming SS—high predictability—patients to a high load of incoming ES—low predictability—patients, who only differ in their level of predictability while holding constant the effects of high numbers of incoming surgical patients. For

hypotheses 4 and 5, we repeat the methods for hypotheses 2 and 3, but under conditions where hospital congestion is high.

Finally, to test whether any changes in LOS due to inventory load affect patient health outcomes, we look at the probability of dying in the hospital or being readmitted within 30 days, given the inventory load on the day before expected discharge. In-hospital mortality and readmission within 30 days are common outcome measures in healthcare that are associated with quality of care (Weingarten, Riedinger et al. 1998; Librero, Peiró et al. 1999). We run a logit model for each outcome variable (in hospital mortality, readmission within 30 days of discharge), controlling for patient gender, age, and the inventory group that characterizes the inventory levels on the day before expected discharge (these groups correspond to the eight survival analysis functions with different ranges of patient inventory, which can be found in Appendix 2).

6. Results

In our sample, we had 283 hospitals with a total of 5.6 million patient admissions. Of these, 126,128 were admissions for hip or knee replacement. The average and median LOS of patients with hip or knee replacement in our sample was 4.46 and 4 days, respectively, and the LOS ranged from 1 to 59 days. Table 1 provides summary statistics for the hip and knee replacement patients.

TABLE 1: SUMMARY STATS HERE

6.1 Main Effects

As Model 1 in Table 2a shows, high congestion (Hypothesis 1) on the day before expected discharge increases the LOS for a "typical" patient with knee/hip replacements by up to 0.8112 days, or 22.8% (t-test=19.8, $p<0.01$) when compared to the baseline case, which we define as a medically identical patient who experiences the same initial hospital conditions, but who has no type of high inventory load (congestion or incoming) on the day prior to discharge. Thus, Hypothesis 1 is supported. For the remainder of the article, assume that when we compare current inpatients, the underlying patients are medically identical, with the demographics of a "typical" patient, and experience all of the same initial hospital conditions; the only differences are the level of congestion and incoming inventory types on the day before expected discharge. Furthermore, the percentage change in LOS is calculated as the change in expected LOS over the expected LOS in the control case for each model. Returning to the role of congestion, we can also see the effect in Figure 2a, which plots LOS of time *t*, in days, on the x-axis, and the probability of "surviving" past time *t* on the y-axis. The survival curve for a patient with high congestion on the day before expected discharge (solid line) is up and to the right of the baseline low congestion curve (dashed line), which means patients who are inpatients with high congestion on the day before their expected discharge have a longer LOS than those with no congestion or other high inventory

load. Looking more closely at the difference in daily probabilities of discharge between patients with and without congestions (see Table 2b), we see that the absolute difference in probability of being discharge on day 3 is 21.1% ($t=23.9$, $p<0.01$) lower for patients who are present during high congestion compared to those with no congestion. While on days 4 through 6, high congestion patients have a higher probability of being discharged on each of these days. These results suggest that current patients who experience high congestion on the day before expected discharge are less likely to be discharged earlier (day 3), but more likely to be discharged later (days 4-6) compared to patients with no congestion, indicating delayed discharges, and thus increasing their expected LOSs.

INSERT FIGURES 2A, 2B, 2C HERE

To isolate the effects of BN pressure (Hypothesis 2), we compare the expected LOSs of patients who are in the hospital when there is a heavy load of incoming low BN pressure patients (EM) to a heavy load of incoming high BN pressure patients (ES, the control case). In this way, we control for any effect that additional incoming patients have on the LOS. In support of H2, Table 2a, Model 2 shows that current hip/knee replacement inpatients who are in a hospital with a heavy incoming load of low BN pressure EM patients on the day before their expected discharge have a LOS that is up to 0.47 days, or 11.7%, longer than patients who are in a hospital with a heavy incoming load of high BN pressure ES patients on the day before expected discharge ($t=9.49$, $p<0.01$). Figure 2b shows the survival curves for these two different groups of hip/knee replacement patients, with the survival curve for high incoming EM patient load being up and to the right of the survival curve for high incoming ES patient load, indicating a longer LOS in the presence of low BN Pressure incoming patients. Furthermore, since both high EM and high ES loads are up and to the right of baseline patients, workers do not appear to speed up with a high ES patients load; instead the change in LOS is due to worker slow-down from a high incoming EM patient load. Looking at the probability of being discharged on days 3-6 (Model 2, Table 2b), we find that a patient who is subjected to a high load of incoming low BN pressure patients on the day before expected discharge is significantly less likely to be discharged on days 3 and 4, but more likely to be discharged on days 5 and 6, suggesting a delayed discharge and longer LOS for this patient.

We follow a similar procedure to determine the effect of predictability on LOS (Hypothesis 3), and compare incoming SS patients, who have high predictability to incoming ES patients, who have low predictability (the control case). As shown in Table 2a, Model 3, we find that, on the day before expected discharge, a heavy load of incoming SS patients, which is associated with high predictability, decreases current hip/knee patients' LOS by up to 0.45 days (10.2%) compared to a high load of incoming ES patients ($t=8.31$, $p<0.01$), supporting H3. In Figure 2c we see that the survival curve for incoming SS patients is down and to the left of the SS curve, so incoming SS patients decrease the LOS of current patients compared to incoming ES patients. At the day level, a current patient with a high number of

predictable incoming patients is significantly more likely to be discharged on day 3, but less likely to be discharged on days 4-6 (see Table 2b, Model 3). The increased probability of being discharged on day 3 indicates that when there are known patients being admitted, workers will reduce processing times for current patients in order to discharge them "early" to make way for the new patients.

TABLES 2a AND 2b ABOUT HERE

6.2 Interaction Effects

Tables 3a, 3b, and 3c show the interaction effects of high congestion with low BN pressure inventory, and high congestion with high predictability inventory on LOS. We follow the same analytical approach as above to measure the effects of incoming inventory with BN pressure (Hypothesis 4) and predictability on current patients (Hypothesis 5), with the adjustment that in all cases, congestion is high. Comparing EM to ES patients under high congestion, we see that high incoming inventory with low BN pressure on the day before expected discharge for a current patient is still associated with an increased LOS for that current patient (Figure 3a). As Table 3a, Model 1 shows, under high congestion conditions, when there are high numbers of incoming patients with low BN pressure on day 3, the LOS of current patients is increased by 0.24 days, or 5.2%, compared to patients who experience incoming patients with high BN pressure and high congestion ($t=8.32$, $p<0.01$). However, the increase in LOS associated with low BN pressure is moderated by high congestion, in support of Hypothesis 4. Low BN pressure without congestion increases the expected LOS of a patient by up to 0.46 days, while low BN pressure with congestion only increases LOS by up to 0.24 days, a reduction in the effect of low BN pressure on LOS of 0.23 days, or 5.2% of the total LOS ($t=4.01$, $p<0.01$) (Table 3b, Model 1). These results suggest that the reduction in LOS from high BN pressure is reduced when there is less slack in the system. Model 1 in Table 3c shows that the probability of being discharged is significantly reduced on days 3-6 when BN pressure is low compared to when BN pressure is in the presence of congestion.

TABLES 3a, 3b, 3c ABOUT HERE

Similarly, we compare the effect on LOS of patients who are present when there is a high load of incoming ES patients to those present during a high load of incoming SS patients under conditions of high congestion (Figure 3b) to obtain the effect of high predictability interacted with high congestion. We find that incoming patients with high levels of predictability do still reduce the LOS of current patients, with a decrease in LOS of up to 0.18 days, or 3.8% ($t=4.85$, $p<0.01$) (See Table 3a, Model 2). As with low BN pressure, the effect of predictability is attenuated by high congestion. The reduction in LOS associated with high predictability is reduced to 0.27 days ($t=4.07$, $p<0.01$) (Model 2, Table 3b), in support of Hypothesis 5. Interestingly, when there is no congestion, high predictability increases the probability of being discharged on day 3 by more than 10%, but lowers the probability on later days, thus reducing the

LOS, while high predictability and high congestion increases the probability of being discharged on both days 3 and 4 (See Table 3c, Model 2). These results imply that predictability still encourages early discharge to make room for the new patients, but more of this discharge occurs on day 4 as opposed to day 3 since there is not enough slack capacity on day 3 due to the high congestion, thus reducing the benefits on LOS of high predictability.

6.3 Patient Outcomes

We also analyze if patient outcome quality is affected by the changes in LOS that result from increased inventory. We compare mortality and readmission rates for all patients and find that there is no association between the rates of mortality and readmission and the level of current or incoming inventory at the p=0.05 significance level. (See Appendix 2).

7. Discussion and Future Research

We have contributed to the field of productivity in operations management by categorizing high load into current in-process inventory (congestion) and incoming inventory, and then further decomposing the latter by its levels of BN pressure and predictability, and quantifying the effect of each on processing times in a healthcare setting. Our results reconcile the perceived discrepancies in the effects of busyness observed in the literature, and allow us to predict a priori how processing times will change with inventory load, as well as their interactions.

Our results support that heavy load plays a significant role in processing times, and the location and characteristics of the inventory load matters in terms of the direction and magnitude of its effect on processing times. We find that high congestion increases LOS by up to 0.81 days, which indicates inefficiency due to overloading of resources. LOS also increases when there is a high incoming inventory load with low BN pressure (up to 0.46 days), consistent with social loafing. Meanwhile, incoming inventory load with high predictability reduces LOS by up to 0.45 days, reflecting workload smoothing, which is enabled by the ability of a worker to plan in advance for a new work assignment by discharging a patient to make room for the incoming one. These results are consistent with the observations in the literature that high load has been shown to both increase and decrease processing times in different settings. By refining the definition of busyness, we are able to predict, a priori, how high inventory load will affect processing times.

Furthermore, we have shown that when high congestion interacts with a high load of predictable incoming inventory, the effects of the predictable incoming inventory on LOS are mitigated, suggesting that much of the speeding up that can be induced by highly predictable inventory requires that workers have some slack capacity. One interesting observation is that with highly predictable incoming patients

and no congestion on the day before expected discharge, there is a shift toward discharging patients currently in the hospital one day earlier than expected (day 3 in this setting). When congestion is added to highly predicable incoming patients, there is still a tendency to discharge current patients early, but due to congestion-related delays, the shift is split across days 3 and 4, suggesting that a worker might be planning ahead, but cannot accomplish all tasks to discharge a patient as quickly because she has more patients to care for. We see a similar dampening effect from high congestion on the impact of LOS when there is a high load of incoming patients with low BN pressure. In this case, the high congestion increases the LOSs with high BN pressure incoming patients more than low BN pressure incoming patients because the benefits of further slowing down are less when there is less slack capacity and there is less chance of deferring work. However, when the load comes from incoming high BN pressure patients, congestion lowers the BN pressure, so there is a benefit to slowing down. Thus, we see a greater increase in LOS when there is high congestion and a high load of high BN pressure incoming patients. These results, combined with the magnitude of high congestion's direct effect on changes in LOS, suggests that high congestion dominates the change in processing times due to high load. However, since predictability and BN pressure indirectly affect congestion via increased or decreased LOS of current patients, all must be considered when making capacity allocation decisions.

In the healthcare setting, all changes in patient LOS are particularly significant due to its fixed reimbursement nature. Recently, there has been an increased push to expand the DRG based payment scheme that is currently used by Medicare. In this system, a treatment, such as a hip or knee replacement, is paid a flat fee (with some adjustment for the location of a hospital and its patient population) for each patient, regardless of actual resources used. Therefore, a hospital is incentivized to fully treat the patient to avoid readmission, but to discharge the patient as soon as possible because any extra time in the hospital incurs costs without additional reimbursement, and consequently, any changes in LOS increases based on patient inventory load have significant financial implications in a hospital setting.

One of the greatest implications for practice is where to redistribute or add resources to improve productivity. In addition to high congestion increasing LOS for current patients, we have found that incoming, unrelated patients from the ED and PACU also affect processing times, and therefore any improvements must take a system-wide approach. For example, in hospitals, the frequent thought is that since EDs are often crowded with patients being treated and/or waiting for hospital beds, which can potentially negatively impact treatment (Gesensway 2011; Batt, Terwiesch et al. 2012), additional capacity (both in terms of space and personnel) should be added to the ED. However, while this may solve temporary patient care and bed problems, our results suggest that it could actually reduce productivity in the long term. Adding additional resources to the ED will further lower the BN pressure induced from the EM incoming patients and will increase the LOS for current patients even more, thus

increasing congestion in the hospital, and resulting in longer waits in the ED. Following the same logic, even more additional resources will end up being put into the ED in the future, resulting in a downward spiral of reduced productivity. Instead, our work suggests a hospital would benefit from adding or allocating additional resources to the inpatient hospital units, and counter intuitively, targeting a lower occupancy level to increase productivity. In this way, the BN pressure of the ED is not reduced, and the negative effects of congestion, both direct and the loss of benefit for highly predictable and high BN pressure incoming patients, are lessened. To further improve productivity, the allocated inpatient hospital resources could include adding a nurse on the hospital floors who is solely responsible for discharges and admissions. By adding this nurse, whose performance measures are tied to efficiently admitting and discharging patient, there would be increased BN pressure on the other nurses to discharge when possible because the admission/discharge nurse would be overseeing their work and know if a patient is ready for discharge. Furthermore, since this nurse is involved in admissions, she would be more aware of the known trends in emergency patient arrivals, thus increasing predictability of incoming EM patients, who are often treated as unpredictable. As an additional step, the hospital, either through the discharge/admission nurse, or through current channels, could improve predictability even more by giving more than 24 hours notice for incoming SS patients. Currently, the standard is 24 hours, and while our data does not directly address the value of additional time, more predictability could further aid in work smoothing to reduce processing times when the incoming patient load is high.

As with all research, there are limitations to this study. First, our dataset only had day level information, so we were unable to measure changes in LOS that did not cross over the midnight hour into the next day. While the exact changes in processing times that result from increased inventory are thus unknown, our results are conservative estimates of the changes in processing times, and do provide an idea of the direction and magnitude of the changes. Second, to isolate the effects of interest, we set incoming inventory levels to 100% or 0% based on being busy or not. We understand that this is an upper bound, but again, it does not affect the direction or order of magnitude of the changes in processing times. Third, since our data is at the hospital level, we do not have occupancy levels for the specific unit within the hospital a patient is being treat in, and the occupancy of these could potentially vary from the overall hospital occupancy. We have made an assumption that the hospital occupancy is highly correlated with the occupancy of any specific unit, and if anything this is a conservative measure of occupancy. Fourth, and related to the previous concern of occupancy, is that we do not have the exact bed capacities of each hospital, nor their staffing levels on each day, so we have had to estimate the occupancy levels. We try to control for differences in capacity on weekends and across quarters of the year, and use the maximum observed as the maximum possible. Fifth, while we have some procedures performed for each patient, we only have major procedures and cannot analyze how changes in inventory affect resource utilization

beyond a bed. It would be interesting to know if the changes in processing times from increased inventory are associated with an increase or decrease in other resources, and if so, what the trade-off is. Overall, we have faced the trade-off of quantity versus granularity of data. Our data has allowed us to quantify trends related to the effects of increased inventory on processing time across hundreds of hospitals. We leave it to future research to analyze the mechanisms for these changes at a more granular level.

FIGURES

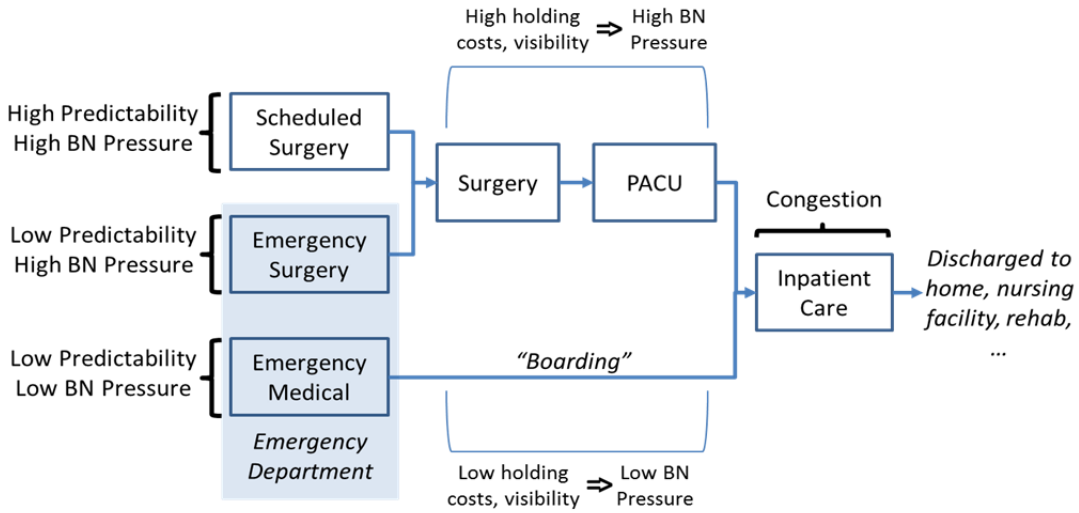

Figure 1. *Paths to Hospital Admission and the Associated Levels of BN pressure and Predictability*

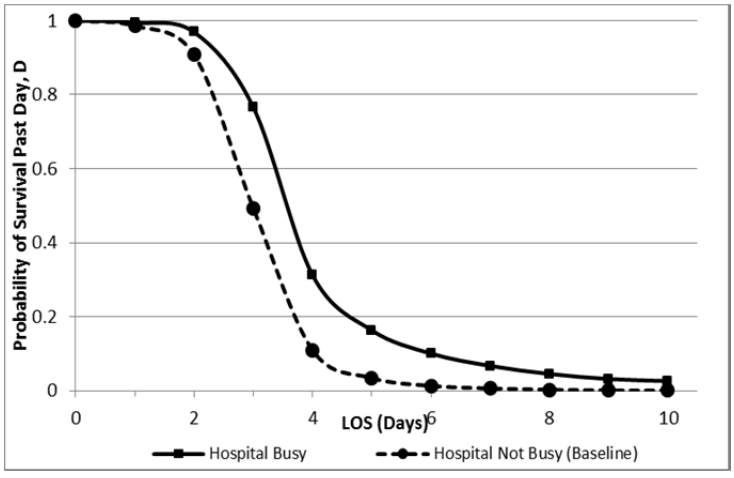

Figure 2a. *Effect of Congestion on LOS*

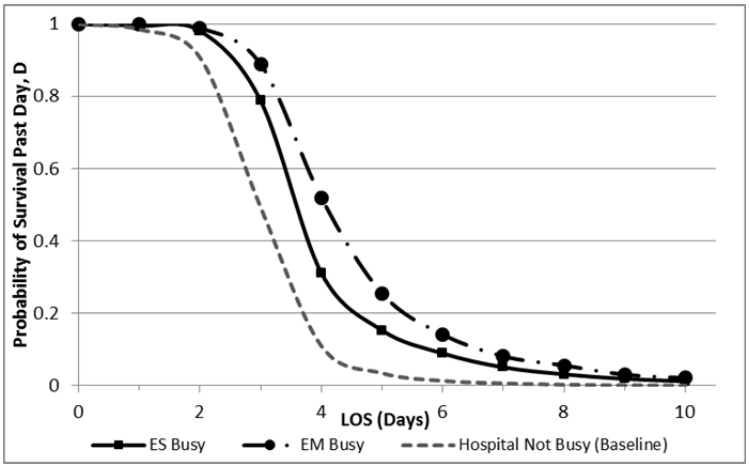

Figure 2b. *Effect of BN pressure on LOS*

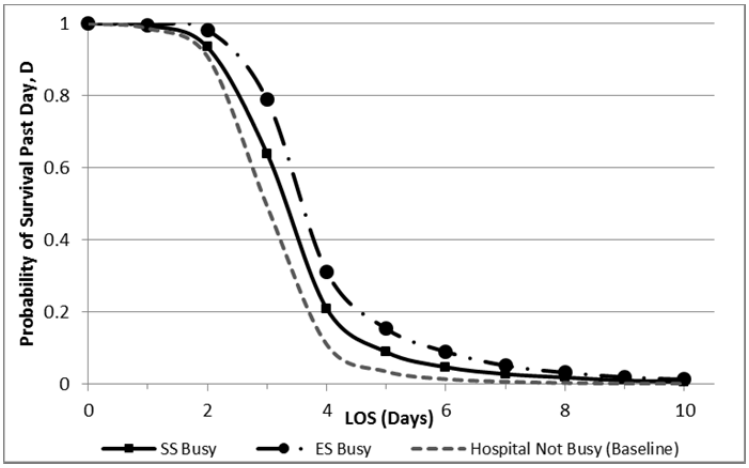

Figure 2c. *Effect of Predictability on LOS*

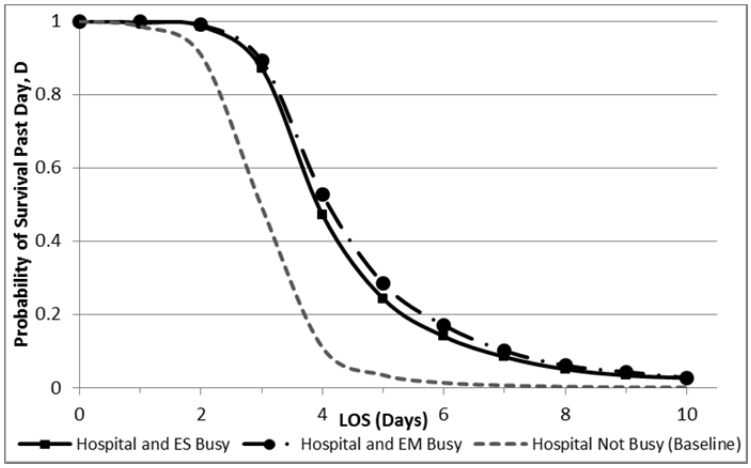

Figure 3a. *Effect of Interaction of BN Pressure and Congestion on LOS*

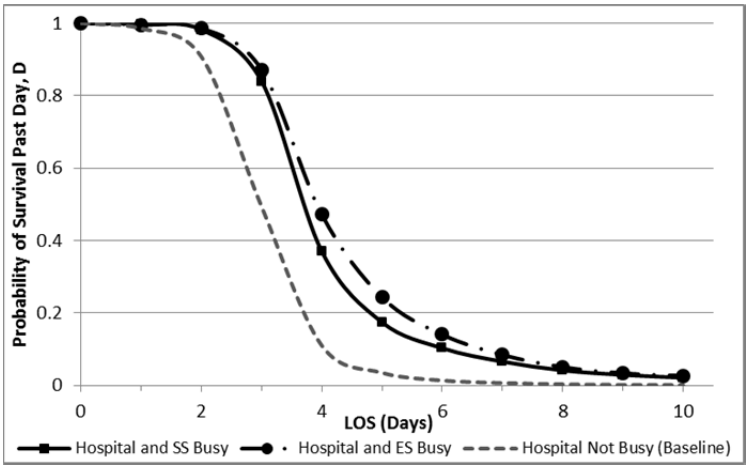

Figure 3b. *Effect of Interaction of Predictability and Congestion on LOS*

TABLES

Hospital Level		**Hip/Knee Replacement Patients**		**Average Occupancy Levels**	
Total Visits	5,609,917	# Visits	126,128	Hospital Occ.	84.3% (10.0%)
# of Hospitals	283	% Female	61.77	SS Occupancy	51.3% (23.2%)
Avg. Visits/Hospital per Year	13,709	Avg. LOS	4.46 (1.61)	ES Occupancy	46.9% (21.3%)
		Avg. Age at Admission	68.11 (11.90)	EM Occupancy	62.2% (16.0%)

Standard deviations in parentheses

Table 1. Summary Statistics

	Model 1	***Model 2***	***Model 3***
	Congestion (n=546)	*BN Pressure (n=546)*	*Predictability (n=546)*
E[LOS] Treatment§ (Days)	4.3649 (0.050)	4.8845 (0.037)	3.9688 (0.049)
E[LOS] Controlǂ (Days)	3.5537 (0.034)	4.4197 (0.051)	4.4197 (0.051)
ΔLOS (days) = Treatment-Control	0.8112** (0.041)	0.4648** (0.049)	-0.4510** (0.054)
% ΔLOS= ΔLOS/E[LOS] Control	22.8%	11.7%	-10.2%
T-value	19.8217	9.4928	8.3081

Standard errors in parentheses; *p<0.05, **p<0.01

§Treatment=High Congestion, Low BN Pressure, High Pred.; ǂControl=Low Congestion, High BN Pressure, Low Pred.

Table 2a. Change in LOS (Main Effects)

	Day 3	**Day 4**	**Day 5**	**Day 6**
Model 1: Congestion (n=546)				
Prob. of Discharge: High Congestion	0.2027 (0.008)	0.4536 (0.008)	0.1490 (0.005)	0.0626 (0.003)
Prob. of Discharge: Low Congestion(Control)	0.4142 (0.012)	0.3838 (0.009)	0.0751 (0.006)	0.0211 (0.002)
Δ in Prob. of Discharge= High Congestion-Control	-0.2114** (0.009)	0.0698** (0.014)	0.0739** (0.005)	0.0414**(0.003)
T-value	23.5372	5.0833	16.0386	13.7535
Model 2: BN pressure (n=546)				
Probability of Discharge: Low BN Pressure	0.0997 (0.004)	0.3693 (0.006)	0.2646 (0.003)	0.1135 (0.002)
Probability of Discharge: High BN Pressure (Control)	0.1932 (0.007)	0.4765 (0.008)	0.1586 (0.005)	0.0637 (0.004)
Δ in Prob. of Discharge= Low BN Pressure-Control	-0.0935** (0.007)	-0.1072** (0.008)	0.1060** (0.007)	0.0498**(0.004)
T-value	13.0087	13.0569	15.8205	12.7457

Model 3: Predictability (n=546)								
Probability of Discharge: High Predictability	0.2949	(0.010)	0.4298	(0.005)	0.1205	(0.007)	0.0419	(0.003)
Probability of Discharge: Low Predictability (Control)	0.1932	(0.007)	0.4765	(0.008)	0.1586	(0.005)	0.0637	(0.004)
Δ in Prob. of Discharge= High Predictability-Control	0.1016**	(0.008)	-0.0466**	(0.009)	-0.0382**	(0.006)	-0.0218**	(0.003)
T-value	12.3703		5.141		6.6238		6.3927	

*p<0.05, **p<0.01

Table 2b. Change in probability of discharge on day, d (Main Effects)

	Model 1	**Model 2**
	BN Pressure w/High Congestion (n=546)	*Predictability w/High Congestion (n=546)*
E[LOS] Treatment§ (Days)	4.9977 (0.030)	4.5780 (0.044)
E[LOS] Controlǂ (Days)	4.7602 (0.029)	4.7602 (0.029)
ΔLOS (days) = Treatment-Control	0.2375** (0.029)	-0.1821** (0.038)
% ΔLOS= ΔLOS/E[LOS] Control	5.2%	-3.8%
T-value	8.3223	4.8450

§Treatment=Low BN Pressure, High Predictability; ǂControl=High BN Pressure ,Low Predictability

*p<0.05, **p<0.01

Table 3a. Change in LOS (Interaction Effects)

	Model 1	***Model 2***
	BN Pressure (n=546)	*Predictability (n=546)*
ΔLOS Interaction Effect§ (Days)	0.2375** (0.029)	-0.1821** (0.038)
ΔLOS Main Effectǂ (Days)	0.4648** (0.049)	-0.4510** (0.054)
Difference in ΔLOS = Interaction-Main	-0.2272** (0.057)	0.2688** (.066)
T-value	4.0098	4.0717

§Interaction Effect=Low Congestion and Low BN Pressure; Low Congestion and High Predictability
ǂMain Effect=High Congestion and Low BN Pressure; High Congestion and High Predictability

*p<0.05, **p<0.01

Table 3b. *Effect of High Congestion on Changes in LOS*

	Day 3	**Day 4**	**Day 5**	**Day 6**
Model 1: BN Pressure w/High Congestion (n=546)				
Probability of Discharge: Low BN Pressure	0.0981 (0.004)	0.3649 (0.005)	0.2424 (0.003)	0.1142 (0.002)
Probability of Discharge: High BN Pressure (Control)	0.1149 (0.004)	0.3996 (0.006)	0.2288 (0.002)	0.1030 (0.002)
Δ in Prob. of Discharge= Low BN Pressure-Baseline	-0.0168**(0.004)	-0.0347** (0.006)	0.0136** (0.002)	0.0112** (0.002)
T-value	4.1799	6.118	5.5744	6.5854
Model 2: Predictability w/High Congestion (n=546)				
Probability of Discharge: High Predictability	0.1427 (0.004)	0.4697 (0.007)	0.1958 (0.003)	0.0715 (0.003)
Probability of Discharge: Low Predictability(Control)	0.1149 (0.004)	0.3996 (0.006)	0.2288 (0.002)	0.1030 (0.002)
Δ in Prob. of Discharge= High Predictability-Control	0.0278** (0.005)	0.0701** (0.006)	-0.0329**(0.003)	-0.0315** (0.002)
T-value	6.209	10.8785	10.0094	13.9547

Standard errors in parentheses; *p<0.05, **p<0.01

Table 3c. *Change in probability of discharge on day, d (Interaction Effects)*

APPENDICES

Appendix 1: Controls

Patient Level	Hospital Level
Gender [Male]	*Year [2008]*
Age at Admission [55]	*Quarter of year [2]*
Race Group	*Day of week of admission [Tuesday]*
White, Black, Hispanic, Asian/Pacific Islander/Aleut, Other *[White]*	*Hospital*
	MRI usage (Day 0, Day 3)[§] *[50%, 50%]*
Admission Type	*ED occupancy* (Day 0, Day 3) *[75%, 75%]*
Scheduled, unscheduled, n/a *[unscheduled]*	*Hospital occupancy* (Day 0) *[86.5%]*
Disposition	*Scheduled Surgical admissions* (Day 0) *[38%]*
Home, Died, Left Against Medical Advice, Other (e.g. Residential care, rehab facility) *[Home]*	*Emergency Surgical admissions* (Day 0) *[44%]*
	Emergency Medical admissions (Day 0) *[62%]*
Payer category	*Severity[ł] of scheduled surgical admissions* (Day 3) *[1.43]*
Medicare, Medi-cal, Private Coverage, Workers' Comp., County Indigent Program, Other Government, Other Indigent, Self Pay, Other [*Medicare]*	*Severity of emergency surgical admissions* (Day 3) *[1.65]*
	Severity of emergency medical admissions (Day 3) *[0.83]*

[§]*A control for relative resource availability*

[ł]*The median DRG weight from CMS for all patients admitted to hospital on day, d*

"Typical" patient value used in model in [ ]

Appendix 2: Survival Function Groups and Quality of Care

Group #	Description
1	Hospital, SS, ES, EM not busy (*Baseline*)
2	High SS patient load; Hospital, ES, EM not busy
3	High ES patient load; Hospital, SS, EM not busy
4	High EM patient load; Hospital, SS, ES not busy
5	High hospital load; SS, ES, EM not busy
6	High hospital and SS patient load; ES, EM not busy
7	High hospital and ES patient load; SS, EM not busy
8	High hospital and EM patient load; SS, ES not busy

	Probability of Dying		Probability of Readmission	
Age	0.1488**	(0.016)	0.0624**	(0.006)
Sex	-0.3184	(0.229)	-0.2659**	(0.080)
Group				
2	0.4321	(0.494)	0.2063	(0.178)
3	0.4619	(0.454)	-0.0563	(0.172)
4	0.3348	(0.649)	0.1590	(0.171)
5	0.7395	(0.394)	0.0526	(0.142)
6	0.5255	(0.438)	0.0352	(0.152)
7	-0.1434	(0.503)	0.1026	(0.129)
8	0.7395	(0.394)	0.1141	(0.150)
Constant	-18.4707**	(1.330)	-9.3156**	(0.476)
n	*126128*		*126128*	
Wald chi2 (11)	*100.14***		*152.36***	
Pseudo R2	*0.1276*		*0.0384*	

Standard errors in parentheses; *p<0.05, **p<0.01

References

Armony, M. and I. Gurvich (2010). "When Promotions Meet Operations: Cross-Selling and Its Effect on Call Center Performance." Manufacturing & Service Operations Management **12**(3): 470-488.

Batt, R. J., C. Terwiesch, et al. (2012). "Docs Under Load: An Empirical Study of State-Dependent Service Rate Mechanisms." Working Paper.

Cachon, G. P. and M. L. Fisher (2000). "Supply Chain Inventory Management and the Value of Shared Information." Management Science **46**(8): 1032-1048.

Chalfin, D. B., S. Trzeciak, et al. (2007). "Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit *." Critical Care Medicine **35**(6): 1477-1483 1410.1097/1401.CCM.0000266585.0000274905.0000266585A.

CMS. (2010). "List of MS-DRGs, Relative Weighting Factors, and Geometric and Arithmetic Mean Length of Stay (FY 2010 Final Rule)." 2012, from http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Acute-Inpatient-Files-for-Download-Items/CMS1247873.html.

Dallery, Y. and S. B. Gershwin (1992). "Manufacturing flow line systems: a review of models and analytical results." Queueing Systems **12**(1): 3-94.

Eddy, D. M. (1984). "Variations in Physician Practice: The Role of Uncertainty." Health Affairs **3**(2): 74-89.

Fisher, M. and A. Raman (1996). "Reducing the Cost of Demand Uncertainty through Accurate Response to Early Sales." Operations Research **44**(1): 87-99.

Fisher, M. L. and C. D. Ittner (1999). "The Impact of Product Variety on Automobile Assembly Operations: Empirical Evidence and Simulation Analysis." Management Science **45**(6): 771-786.

George, J. M. and J. M. Harrison (2001). "Dynamic Control of a Queue with Adjustable Service Rate." Operations Research **49**(5): 720-731.

Gesensway, D. (2011) "Admissions from the ED: Are patients leaving too fast or too slow " Today's Hospitalist.

Green, L., S. Savin, et al. (2011). ""Nursevendor Problem": Personnel Staffing in the Presence of Endogenous Absenteeism." Working Paper(h).

Green, L. V. (2002/2003). "How Many Hospital Beds?" Inquiry **39**: 400-412.

Green, L. V. and V. Nguyen (2001). "Strategies for Cutting Hospital Beds: The Impact on Patient Service." Health Services Research **36**(2): 421-442.

Ha, A. Y. (1998). "Incentive-Compatible Pricing for a Service Facility with Joint Production and Congestion Externalities." Management Science **44**(12): 1623-1636.

Hoetker, G. (2007). "The Use of Logit and Probit Models in Strategic Management Research: Critical Issues." Strategic Management Journal **28**: 331-343.

Hopp, W. J., S. M. R. Iravani, et al. (2007). "Operations Systems with Discretionary Task Completion." Management Science **53**(1): 61-77.

Jenkins, S. P. (2004). Survival Analysis. Unpublished Manuscript. Colchester, UK, Institute for Social and Economic Research, University of Essex.

KC, D. and C. Terwiesch (2009). "Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations." Management Science **55**(9): 1486-1498.

KC, D. S. and C. Terwiesch (2012). "An Econometric Analysis of Patient Flows in the Cardiac Intensive Care Unit." Manufacturing & Service Operations Management **14**(1): 50-65.

Kuntz, L., R. Mennicken, et al. (2012). "Stress on the Ward: Evidence of Safety Tipping Points in Hospitals." Working Paper.

Latané, B., K. Williams, et al. (1979). "Many hands make light the work: The causes and consequences of social loafing." Journal of Personality and Social Psychology **37**(6): 822-832.

Librero, J., S. Peiró, et al. (1999). "Chronic Comorbidity and Outcomes of Hospital Care: Length of Stay, Mortality, and Readmission at 30 and 365 Days." Journal of Clinical Epidemiology **52**(3): 171-179.

Little, J. D. C. (1961). "A Proof for the Queuing Formula: L = λW." Operations Research **9**(3): 383-397.

Litvak, E., Ed. (2010). Managing Patient Flow in Hospitals: Strategies and Solutions. Strategies for managing patient flow. Pg 57-74. Oakbrook Terrace, IL, Joint Commision Resources.

Loch, C. H. and C. Terwiesch (2005). "Rush and Be Wrong or Wait and Be Late? A Model of Information in Collaborative Processes." Production and Operations Management **14**(3): 331-343.

Macario, A., F. Dexter, et al. (2001). "Hospital Profitability per Hour of Operating Room Time Can Vary Among Surgeons." Anesthesia & Analgesia **93**(3): 669-675.

Mas and Moretti (2007). "Peers at Work."

Mas, A. and E. Moretti (2007). "Peers at Work." American Economic Review **99**(1): 112-145.

Matthews, A. L., C. M. Harvey, et al. (2002). "Emergency Physician to Admitting Physician Handovers: An Exploratory Study." Proceedings of the Human Factors and Ergonomics Society Annual Meeting **46**(16): 1511-1515.

McLeod, P. J., R. M. Tamblyn, et al. (1997). "Use of standardized patients to assess between-physician variations in resource utilization." The Journal of the American Medical Association **278**(14): 1164-1168.

Oliva, R. and J. D. Sterman (2001). "Cutting Corners and Working Overtime: Quality Erosion in the Service Industry." Management Science **47**(7): 894-914.

Powell, A., S. Savin, et al. (2011). "Physician workload and hospital reimbursement: Overworked servers generate lower income." Working Paper.

Powell, S. G. and K. L. Schultz (2004). "Throughput in Serial Lines with State-Dependent Behavior." Management Science **50**(8): 1095-1105.

Roberts, R. R., P. W. Frutos, et al. (1999). "Distribution of Variable vs Fixed Costs of Hospital Care." JAMA: The Journal of the American Medical Association **281**(7): 644-649.

Schultz, K. L., D. C. Juran, et al. (1999). "The Effects of Low Inventory on the Development of Productivity Norms." Management Science **45**(12): 1664-1678.

Schultz, K. L., D. C. Juran, et al. (1998). "Modeling and Worker Motivation in JIT Production Systems." Management Science **44**(12): 1595-1607.

Shapiro, R. D. (1996). "National Cranberry Cooperative." Harvard Business School Case **#688-122**.

Spear, S. and H. K. Bowen (1999). "Decoding the DNA of the Toyota Production System." Harvard Business Review **September-October**: 96-106.

Stidham, S. and R. R. Weber (1989). "Monotonic and Insensitive Optimal Policies for Control of Queues with Undiscounted Costs." Operations Research **37**(4): 611-625.

Tan, T. F. and S. Netessine (2012). "When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity." Working Paper.

Tucker, A. L. and S. J. Spear (2006). "Operational Failures and Interruptions in Hospital Nursing." Health Services Research **41**(3p1): 643-662.

Waddle, J. P., A. S. Evers, et al. (1998). "Postanesthesia care unit length of stay: quantifying and assessing dependent factors." Anesthesia & Analgesia **87**(3): 628-633.

Weingarten, S., M. Riedinger, et al. (1998). "Can practice guidelines safely reduce hospital length of stay? Results from a multicenter interventional study." The American journal of medicine **105**(1): 33-40.

Ziser, A., M. Alkobi, et al. (2002). "The postanaesthesia care unit as a temporary admission location due to intensive care and ward overflow†‡." British Journal of Anaesthesia **88**(4): 577-579.